**Name: Arif Pasha**

Sr Data Engineer/ Big Data | 630-384-9467| Email: pash.arif007@gmail.com

**PROFESSIONAL SUMMARY:**

* Over 12+ years of IT experience as a Senior Big Data Engineer and Data Analyst, specializing in SQL, Python/Scala, Spark API, and UNIX shell scripting for data processing.
* Proficient in Big Data/Hadoop Ecosystem, including Apache Hive, Spark, MapReduce, Apache Kafka, Sqoop, Oozie, Zookeeper, HDFS, and YARN.
* Skilled in using libraries like NumPy, SciPy, Requests, Report Lab, Pytables, cv2, HTTPLib, Urllib, Beautiful Soup, and Pandas during the development life cycle.
* Experienced with various Hadoop Distributions such as Cloudera CDH, Amazon Elastic MapReduce, and Hortonworks Data Platform.
* Expertise in optimizing and tuning Spark components for performance and cost efficiency.
* Configured Spark connections for batch and real-time data processing using HDFS and in-memory Spark Data frame API.
* Extensive use of Spark Data frame API on the Cloudera Platform for analytics and data manipulation. • Proficient in integrating NoSQL databases like DynamoDB, Cassandra, HBase, and Cosmos DB.
* Skilled in writing scripts utilizing Spark data frames APIs for data processing and manipulation, as well as Spark Streaming for real-time data processing from Apache Kafka.
* Strong ability to interact with HDFS using HiveQL for ad-hoc data extraction and analysis, along with writing Hive User Defined Functions (UDF).
* Expertise in partitioning and bucketing techniques for optimizing query performance.
* Proficient in translating SQL/Hive queries to Spark transformations in Data Frame and Data Set APIs.
* Knowledgeable in NoSQL databases such as MongoDB, HBase, and Apache Cassandra.
* Hands-on experience with SQL and NoSQL databases like Oracle, MongoDB, Cassandra, DynamoDB, and CouchDB.
* Familiar with version control systems like Git, GitHub, CVS, and SVN. • Skilled in writing maintainable and clean code for reuse and longevity.
* Experience in developing web applications following MVC architecture using Django, Flask, and Python.
* Conducted ad-hoc Data Analysis on large datasets to provide data-driven insights to business leaders.
* Proficient in Data Preprocessing, Data Analysis, and Machine Learning for gaining insights from structured and semi-structured data.
* Skilled in writing SQL queries, Stored Procedures, functions, packages, tables, views, and triggers on relational databases like Oracle, DB2, MySQL, PostgreSQL, and MS SQL Server.
* Experienced in using Docker and Ansible for automation and deployment. • Familiar with containerization and orchestration services like Docker and Kubernetes.
* Excellent communication, interpersonal, problem-solving skills, and team player.
* Successfully implemented Proof of Concept projects in the Big Data/Hadoop ecosystem to reduce processing latency.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Hadoop/Big Data Technologies | **Hadoop, Map Reduce, Sqoop, Hive, Oozie, Spark, Zookeeper and Cloudera Manager, Kafka, Flume.** |
| ETL Tools | **SAP BODS, Teradata** |
| NO SQL Database | **HBase, Cassandra, Dynamo DB, Mongo DB.** |
| Monitoring and Reporting | **Tableau, Custom shell scripts, Apache Airflow** |
| Hadoop Distribution | **Horton Works, Cloudera, Amazon EMR** |
| Build Tools | **Maven** |
| Programming & Scripting | **Python, Scala, JAVA, SQL, Shell Scripting, C, C++** |
| Databases | **Oracle, MY SQL, Teradata** |
| Version Control | **GIT, GitHub** |
| Operating Systems | **Linux, Unix, Mac OS-X, CentOS, Windows 10, Windows 8, Windows 7** |
| Cloud Computing | **AWS, Oracle, Azure** |
| Data base modelling | **ER modelling, dimension modelling, Start schema modelling, Snowflake modelling** |
| Visualization Reporting | **Tableau, ggplot, matplotlib and PowerBI** |
| Machine Learning Techniques | **Linear and Logistic Regression, Clustering, Random Forest, NLP** |

**PROFESSIONAL EXPERIENCE:**

**Client: Capstone Logistics, Chicago, IL June 2020- Present**

**Role: Sr. Data Engineer/Architect**

**Responsibilities:**

* Vast experience in utilizing the **AWS cloud platform, including EC2, S3, EMR, Redshift, Lambda, and Glue.**
* **Designed, implemented, and managed end-to-end data pipelines using ADF** for efficient data **integration, transformation, and movement** across diverse data sources.
* **Leveraged ADF to enable the smooth movement of data** between various data stores, supporting a hybrid cloud environment.
* **Improved data processing efficiency by optimizing and tuning ADF** pipelines, resulting in reduced processing times and increased overall system performance.
* **Implemented data quality checks within ADF pipelines**, enhancing the reliability and accuracy of the integrated data.
* Collaborated with **data scientists, analysts, and business stakeholders to understand data requirements** and deliver tailored solutions using **ADF**.
* Implemented **security measures within ADF to ensure data encryption**, authentication, and compliance with industry standards.
* Led the implementation **of Azure Databricks** to create **scalable and collaborative data** processing workflows.
* Developed and executed complex data transformation and analytics workflows using **Spark SQL and PySpark within Azure Databricks.**
* Integrated machine learning models into data processing pipelines using MLlib and MLflow within **Azure Databricks for predictive analytics.**
* Managed **and optimized large volumes of structured and semi-structured data** **within Azure Synapse Analytics,** ensuring efficient data storage and retrieval.
* Leveraged **the MPP architecture of Synapse Analytics** for distributed processing, significantly **improving query performance** on **large datasets**.
* Developed and **executed complex SQL queries and analytics workflows within Synapse Analyti**cs, enabling data analysts and scientists to derive valuable insights.
* Qlik Compose automates the entire ETL process, from data extraction to loading and transformation. This saves data engineers time on manual coding and allows them to focus on other high-priority tasks.
* It enables the creation of complex transformations through an intuitive interface, reducing the need for extensive manual coding.
* **Integrated Azure Synapse Analytics with big data technologies,** allowing seamless analysis of both relational and non-relational data within a unified environment.
* **Data engineers can continuously replicate**
* **data from source to target systems in real time with minimal latency. This ensures that analytics teams have access to the most up-to-date data for decision-making.**
* **When migrating data between environments (e.g., from on-premises to the cloud), Qlik Replicate ensures zero downtime replication, minimizing disruption to business operations during data migrations.**
* **Qlik Replicate uses CDC technology to track and capture changes made to source data in real time, ensuring that only the changes are replicated, reducing resource consumption and optimizing performance for data engineers.**
* **Collaborated with data science teams** to implement advanced analytics and machine learning workflows **within Synapse Analytics, enabling predictive modeling** and insights generation.
* Proficiently handled Spark components such as **Spark RDD, Data frame API, Data set API, Spark SQL, Data Source API, and Spark Streaming.**
* Developed Spark-based applications using **Python and Apache Spark** for efficient data processing, handling data from various **RDBMS and Streaming sources**.
* Successfully migrated code to work with the **Spark execution engine, resulting in improved performance and optimization of existing Hadoop applications**.
* **Utilized Spark Streaming APIs for** real-time data transformation and loading.
* Designed a data model to capture real-time data from Kafka and persist it in **Cassandra.**
* Developed a Python Kafka consumer API for data consumption from topics. • **Skillfully parsed Extensible Markup Language (XML)** messages using Kafka and processed XML files to capture real-time User Interface (UI) updates.
* Effectively migrated on-premises applications to the **AWS cloud platform, employing AWS EMR to maintain the Hadoop cluster.**
* **Qlik Compose supports secure data pipelines, offering fine-grained access control and encryption, which is essential for ensuring that sensitive data is protected throughout the pipeline.**
* Loaded data into S3 buckets using **PySpark** and **AWS Glue**, applying Elasticsearch to filter data stored **in S3 buckets** before loading it into Hive External tables.
* Configured a **snow pipe** to facilitate data extraction from S3 buckets into Snowflake tables.
* **Staged API or Kafka data** (in JSON file format) **into Snowflake DB, flattening it for various functional services.**
* Managed incoming data within **Snowflake's staging area.**
* **Worked with Amazon Redshift** to migrate on-premises data warehouses to a unified Data warehouse.
* Developed **Kibana** Dashboards based on log stash data and integrated source and target systems into **Elasticsearch for near real-time log analytics and End-to-End transaction** monitoring.
* Implemented AWS Step Functions to automate and **orchestrate Amazon SageMaker tasks**, such as data publishing **to S3, ML model training, and deployment for prediction**.
* Demonstrated a strong understanding of **Apache Cassandra architecture,** including replication **strategy, gossip, snitches**, and more.
* Played a key role in designing columnar families in **Cassandra, ingesting data from RDBMS, performing data transformations, and exporting data to Cassandra as per business requirements.**
* Proficient in creating data models for client transactional logs, **analyzing data from Cassandra**.
* Utilized **the Spark Data Cassandra c**onnector to seamlessly load data to and from Cassandra.
* Experienced in working with **Parquet, Avro, and JSON file formats, and developed UDFs in Hive**.
* **Developed Sqoop and Kafka jobs** to efficiently load data from RDBMS and External Systems into **HDFS.**
* **Created Oozie coordinators to schedule Hive Scripts and create Data Pipelines.**
* Authored **several MapReduce jobs using Pyspark** and NumPy, leveraging Jenkins for Continuous integration.
* Conducted on-cluster testing **of HDFS, Hive, Pig, and MapReduce** to ensure smooth access for new users.
* Ensured continuous monitoring and management of the Hadoop cluster through Cloudera Manager.
* Worked with Business Analyst to understand the user requirements, layout, and look and feel of the application to be developed.
* O Data blending implemented on SQL Server database and generated interactive dashboards.
* Created ad-hoc reports to users in Tableau by connecting various data sources.
* O Preparing dashboards using calculated fields, parameters, calculations, groups, sets and hierarchies in Tableau.
* O Created sheet selector to accommodate multiple chart types in a single dashboard by using parameters.
* environment and interface correctly with other production systems.
* • Made complex chart like Spider chart and funnel chart to better visualize the data.
* • Generated Dashboards with Quick filters, Parameters and sets to handle views more efficiently.
* Environment: **Spark, Spark SQL, AWS EMR, MapR, HDFS, Hive, Apache Kafka, Sqoop, Python, PySpark, Shell scripting, Linux, MySQL, Oracle Enterprise DB, Eclipse, Jenkins, Oracle, Git, Tableau, MySQL, SOAP, Cassandra, Agile Methodologies, Redshift, Avro, SageMaker.**

**Client: DaVita, San Francisco, CA May 2018 – June 2020**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Developed **Spark** scripts using **Scala** and **PySpark** to meet project requirements.
* Engaged in Spark code development using **Scala and Spark-SQL** to enable faster testing and data processing, optimizing it **through Spark Context, Spark-SQL, PairRDDs, and Spark Yarn.**
* **Managed and maintained Data Pipelines, triggering them and DataBricks Jobs using Apache Lambda.**
* Created roles and **policies using IAM** and monitored Data Pipelines using **AWS CloudWatch.**
* Deployed **Databricks** jobs and created **Data** **Pipelines** using **Jenkins**.
* Loaded data from Informatica server to **HDFS on the EMR service using Sqoop.**
* Generated reports using visualization tools like **Tableau and QlikView.**
* Wrote SQL queries for large data warehousing **applications in Teradata.**
* Set up and maintained data pipelines using workflow management tools like **Apache Airflow.**
* Created data models and implemented **SQL performance tuning techn**iques, as well as schema creation in **Data Catalog.**
* Established and managed **data pipelines, created lambda codes**, checked logs in CloudFormation in AWS, and developed shell scripts in Python or Bash.
* Implemented **data extraction, data modeling, and data manipulation** techniques for large datasets in the petabyte range.
* Qlik Compose offers an end-to-end solution for data ingestion, transformation, and loading into data warehouses or lakes, allowing data engineers to manage the entire data pipeline lifecycle seamlessly.
* Qlik Compose supports real-time data integration and synchronization. This is essential for data engineers who need to ensure data is up-to-date in data lakes or warehouses, especially in dynamic environments where decisions depend on timely data.
* Qlik Replicate supports seamless integration with major cloud platforms such as AWS, Azure, and Google Cloud, enabling data engineers to replicate data into cloud data warehouses and lakes (e.g., Snowflake, Redshift, Google BigQuery).
* For data engineers working in hybrid environments (combination of on-premises and cloud systems), Qlik Replicate supports the replication of data across these different infrastructures without disruption.
* **Created S3 buckets**, managed policies for **S3 buckets**, and utilized S3 bucket and Glacier for storage, including AWS backups and data storage using DynamoDB.
* Created graphical representations using AWS **QuickSight** and crafted queries for Snowflake Databases.
* **Created and maintained** issue log where business users enter their feedback by working on dashboard in test environment.
* **Implemented** the feedback and incorporated the necessary changes.
* Performed Parallel Testing or **Production** Testing, which ensures that the new system will perform correctly in a production
* Wrote **optimized Python code for retrieving data from REST APIs.**
* Environment: **Spark, Spark SQL, Databricks, AWS EMR, S3, AWS Glacier, AWS CloudWatch, HDFS, Hive, Apache Kafka, Sqoop, Python, PySpark, Shell scripting, Linux, Apache Airflow, Jenkins, Oracle, Git, Tableau, DynamoDB, Teradata, Tableau, QlikView, REST API**.

**Client: Ross Stores Inc., Dublin, CA Sept 2017 – Mar 2018**

**Role: Data Engineer**

**Responsibilities:**

* Proficient in Azure cloud platform operations, including **DataLake, DataBricks**, **HDInsight, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH, and Data Storage Explorer.**
* Designed and executed data pipelines on **DataLake using DataBricks** and **Apache Airflow**.
* Empowered data scientists and analysts to work on machine learning solutions effectively.
* Leveraged **Azure Data Factory** to seamlessly integrate both on-premises data sources (MySQL and Cassandra) and cloud-based resources (**Blob storage, Azure SQL DB), implementing necessary transformations for loading into Azure Synapse**.
* **Configured Spark Streamin**g to receive real-time log data from Apache Flume and stored the streaming data in **Azure tables** using Spark.
* **Utilized DataLake** for storage and conducted various processing and analytics tasks on semi-structured and structured data.
* Ingested data into **Blob Storage** and processed it using **Databricks**, involving the development of **Spark Scala scripts and UDFs** for efficient transformations on large datasets.
* Employed **Spark streaming API** to ingest data from diverse sources.
* **Optimized** existing code to enhance cluster performance.
* Utilized **Spark DataFrames** to create different datasets and applied business transformations and data cleansing operations within the **Databricks platform**.
* Developed efficient, reusable, and clean **Python scripts** for building ETL pipelines, creating Direct Acyclic Graph (DAGs) workflows using **Airflow and Apache NiFi.**
* Extensively used **Kubernetes** to manage batch workloads essential for feeding, analytics, and machine learning applications.
* Monitored **Spark clusters** using Log Analytics and **Ambari Web UI**, transitioning log storage from Cassandra to **Azure SQL Datawarehouse** to enhance query performance.
* Effectively managed resources and scheduling across the cluster through **Azure Kubernetes Service (AKS),** enabling the creation, configuration, and management of virtual machines within the cluster.
* Leveraged **Azure DevOps** and **VSTS (Visual Studio Team Services) for CI/CD**, employed Active Directory for **authentication, and implemented Apache Range** for authorization.
* Developed Scala applications for processing large datasets through MapReduce jobs, compiling programs into bytecode for **JVM-based data processing.**
* Proficiently **performed memory tuning, adjusted batch interval times, and optimized parallelism levels** to enhance processing efficiency.
* Utilized data for creating interactive **Power BI** dashboards and reports tailored to meet specific business requirements.
* **Published Workbooks** by creating user filters so that only appropriate teams can view it.
* Analyzed the source data and handled efficiently by modifying the data types. Used excel sheet, flat files, CSV files to generated Tableau **ad-hoc reports**.
* Environment: **Azure HDInsight, DataBricks (ADBX), DataLake (ADLS), CosmosDB, MySQL, Snowflake, MongoDB, Teradata, Ambari, Flume, VSTS, Tableau, PowerBI, Azure DevOps, Ranger, Azure AD, Git, Blob Storage, Data Factory, Data Storage Explorer, Scala, Hadoop 2.x (HDFS, MapReduce, Yarn), Spark v2.0.2, Airflow, Hive, Sqoop, HBase.**

**Client: United Health Group,** [**Minnetonka, MN**](https://www.google.com/search?sca_esv=568517199&rlz=1C1CHBD_enUS1023US1023&sxsrf=AM9HkKkqtWoZ0THh66k-DNfuvQmpqgaiNQ:1695738054161&q=Minnetonka,+Minnesota&stick=H4sIAAAAAAAAAONgVuLQz9U3KDdMSnnEaMwt8PLHPWEprUlrTl5jVOHiCs7IL3fNK8ksqRQS42KDsnikuLjgmngWsYr6ZublpZbk52Un6iiA2cX5JYkA67JB8loAAAA&sa=X&ved=2ahUKEwjotubCvMiBAxXGJTQIHSbBA9AQzIcDKAB6BAghEAE) **May 2016 – Aug 2017**

**Role: Data Engineering Analyst**

**Responsibilities:**

* Enterprise Manufacturing domain **Data Warehouse** project of migrating existing **Data Marts** to an integrated place to exploit the advantage of corporate wide **data warehouse**. Involved redeveloping existing **Data Marts** and adding new subject areas to existing ones.
* Developed automation **Shell scripts** in **Linux** system for processing of manufacturing data from **Blob storage** API (spreadsheet data) and brought turnaround time from 20 minutes to under 10 seconds.
* Designed, Developed, and delivered the jobs and transformations over the data to enrich the data and progressively facilitate the consumption in the **Enterprise** **Data Lake**.
* Performed extensive validation using hive queries on **Data Lake** external tables after loads to check overall data quality in terms of correctness, partitions, performance.
* Performed **EDA** on raw data and identified anomalies. Created **Curation layer** after transforming tables using python data manipulation modules for ML usage.
* Implemented & maintained an end-to-end machine learning based flask application for steel manufacturing client for current steel production application
* Developed **Data pipelines** in **OCI** to extract, transform and load from different sources into **Oracle Data Lake**.
* Upgraded existing loading processing by integrating **Sqoop** functionality and Crontab. Also, presented potential dashboard changes to business in tableau.
* Automated weekly data load from server to data lake using **file transfer protocol(ftp), regular expressions, pandas,** and **shell scripting**
* **Environment:** Oracle Cloud Infrastructure, Python, Apache Spark, Hadoop ecosystem (MapReduce, HDFS, Hive), OBIEE, Linux, Flask, WinSCP, Tableau.

**Client: Bank of America,** [**Charlotte, NC**](https://www.google.com/search?sca_esv=568517199&rlz=1C1CHBD_enUS1023US1023&sxsrf=AM9HkKlnQeJu2NAipFi5dzYABqKSlNejLw:1695738159116&q=Charlotte,+North+Carolina&stick=H4sIAAAAAAAAAONgVuLQz9U3SCtOsnjEaMwt8PLHPWEprUlrTl5jVOHiCs7IL3fNK8ksqRQS42KDsnikuLjgmngWsUo6ZyQW5eSXlKTqKPjlF5VkKDgnFuXnZOYlAgCZ8-lxXgAAAA&sa=X&ved=2ahUKEwjfq-z0vMiBAxXrhIkEHbKqA7cQzIcDKAB6BAgSEAE) **Jan 2013 – Apr 2016**

**Role: ETL/SQL Developer**

**Responsibilities:**

* BODS Consultant in four full life cycle **data migration** implementation projects (Waterfall methodology).
* A data migration project with Business goal to extract, transform and load the data into SAP ECC.
* Developed the code in BODS and HANA Studio to migrate the data of complex object like Pricing Condition, Sales Order, Outputs, Sales Contract and Order Template etc.
* Developed objects based on client requirements using **Workflows**, **Data flows**, **Transformations**, **Global variables**, **Script**, Database data-store Objects (**Calculation views**,

Tables, **Template tables**), and flat files.

* Used different transformations like **Query**, Table Comparison, **Map operation**, **Validation,**

**Case, Row Generation, Pivot, Reverse Pivot, Merge, Key generation, Data transfer, SQL Transform**.

* Experienced in job **debugging**, job check in and check out to central repository and labeling
* Loading the data into target system via **LSMW, IDOC** or Custom Load tools.
* Executed the Jobs for respective environments (Development, Quality, and production
* systems) in Sap data service management console.
* **IDOC** implementation in **BODS**.
* Created Technical Design Documents (**TDD**s) which can track the changes done in the code.
* Worked on HP Application Life cycle management tool for defect analysis and resolution.
* Involved into multiple Innovation and Automation as well as in HANA architecture.
* **Environment:** SAP BODS, SAP HANA Studio, SAP ECC Server, SAP ABAP, HP Application Lifecycle Management Tool.

Education:

Masters in Software Engineering 2010

University of Illinois