**Sharfuddin Syed**

**Sr. Azure Data Engineer**

**Email:** **syedsharfuddin332@gmail.com**

**Ph. No: 940-487-7420**

**Professional Summary**

* Overall **7 Years** of experience in Software industry and **Azure** **data** **engineer**. Experienced in importing real time streaming logs and aggregating the **data** to **HDFS** using **Kafka** and Flume.
* Experience in building **data** pipelines using **Azure** **Data** factory, **Azure** **data** bricks and loading **data** to **Azure** **Data** Lake, **Azure** **SQL** **Database**, **Azure** **SQL** **Data** warehouse and controlling and granting **database** access.
* Experience on Migrating **SQL** **database** to **Azure** **Data** Lake, **Azure** **data** lake Analytics, **Azure** **SQL** **Database**, **Data** Bricks and **Azure** **SQL** **Data** warehouse and controlling and granting **database** access and Migrating On premise databases to **Azure** **Data** lake store using **Azure** **Data** factory. Strong skills in visualization tools Power BI, Confidential Excel - formulas, Pivot Tables, Charts and DAX Commands.
* Experience in Developing **Spark** applications using **Spark** - **SQL** 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.
* Good understanding of **Spark** Architecture including **Spark** Core, **Spark** **SQL**, **Data** Frames, **Spark** **Streaming, Driver Node, Worker Node, Stages,** Executors and Tasks. In-depth knowledge of **Snowflake Database**, Schema, Table structures and **Snowflake cloud** technology
* Good understanding of **Big** **Data** **Hadoop** and Yarn architecture along with various **Hadoop** Demons such as Job Tracker, Task Tracker, Name Node, **Data** Node, Resource/Cluster Manager, and **Kafka**. Experience in analyzing **data** using HiveQL, and MapReduce Programs.
* Experience in **Database** Design and development with Business Intelligence using **SQL** Server 2014/2016, Integration **Services** (SSIS), DTS Packages, **SQL** Server Analysis **Services** (SSAS), DAX, OLAP Cubes, Star Schema and **Snowflake** Schema.
* Experienced in ingesting **data** into **HDFS** from various Relational databases like MYSQL, Oracle, DB2, Teradata, and Postgres using **Sqoop**.
* Well versed with various **Hadoop** distributions which include Cloudera (CDH), Hortonworks (HDP), and **Azure** HD Insight.
* 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. Experience working on NoSQL Databases like HBase, Cassandra and MongoDB.
* Experience in **Python**, **Scala**, shell scripting, and **Spark**. Hands on Experience in performing Host based online SAN migrations.
* Experience on **ETL** methodology for supporting **Data** Extraction, transformations and loading processing using **Hadoop**.
* Automated resulting scripts and workflow using Apache airflow and shell scripting to ensure daily execution in production
* Experience 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.
* Good understanding of **Spark** Architecture, MPP Architecture, including **Spark** Core, **Spark** **SQL**, **Data** Frames, **Spark** Streaming, Driver Node, Worker Node, Stages, Executors and Tasks. Deep Knowledge in ServiceNow tool, Creating New change requests.
* Experience in Developing **Spark** applications using **Spark** - **SQL** 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.
* Working as **Cloud** Administrator on Microsoft **Azure**, involved in configuring virtual machines, storage accounts, resource groups.
* Good understanding of **Spark** Architecture including **Spark** Core, **Spark** **SQL**, **Data** Frames, **Spark** Streaming, Driver Node, Worker Node, Stages, Executors and Tasks. Experience with MS **SQL** Server Integration **Services** (SSIS), T-**SQL** skills, stored procedures, triggers.
* Design and develop **Spark** applications using Pyspark and **Spark**-**SQL** for **data** extraction, transformation and aggregation from multiple file formats for analyzing & transforming the **data** to uncover insights into the customer usage patterns.
* Good understanding of **Big** **Data** **Hadoop** and Yarn architecture along with various **Hadoop** Demons such as Job Tracker, Task Tracker, Name Node, **Data** Node, Resource/Cluster Manager, and **Kafka** (distributed stream-processing). **Azure** **Data** Factory (ADF), Integration Run Time (IR), File System **Data** Ingestion, Relational **Data** Ingestion.
* Experience, specialized in **Azure** **Data** Engineering, **Big** **Data** Ecosystem- **Data** Acquisition, Ingestion, Modelling, Storage Analysis, Integration, **Data** Processing. Experienced working with **snowflake** and has ingested **data** into the **snowflake cloud** **data** warehouse using Snow pipe.
* Extensively used **Kafka** for creating partition, created **Kafka** topics, and written custom partitioner classes using clusters in **Kafka**.
* Extensively used **cloud**-native deployment of confluent **Kafka** using Kubernetes, Docker, and related technologies.
* Proficient in designing, implementing, and managing confluent **Kafka** solutions for high-volume, real time **data** streaming and processing applications. Extensive experience with **Python** coding using arithmetic operations such as pandas, NumPy.
* Extensive Experience in working with **Azure** Databricks, Synapse Analytics, **Azure** **Data** Factory, Stream Analytics. **Azure** Analysis **Services**, **Data** Lake, **Azure** Storage, **Azure** **SQL** **Database**, **SQL** **Data** Warehouse, **Azure** Cosmos DB.
* Expertise in working with **Azure** **services** like HDInsight, Application Insights, **Azure** Monitoring, **Azure** AD, Function apps, Logic apps, Event Hubs, Iot hubs, Storage Explorer, Key Vault. Experience with Testing Map Reduce programs using MR Unit, Junit and Easy Mock.
* Strong working experience with **SQL** and NoSQL databases (**Azure** Cosmos DB, MongoDB, HBase, Cassandra), **data** modelling, tuning, disaster recovery, backup and creating **data** pipelines. Working with Iterative Agile programming methodologies, Waterfall, and Scrum methodologies.
* Proficient in handling and ingesting terabytes of Streaming **data** (**Kafka**, Event Hub, IOT Hub, Kinesis, **Spark** streaming, Storm), Batch **Data**, Automation and Scheduling (Oozie). Have extensive experience in creating pipeline jobs, schedule triggers using **Azure** **data** factory.
* Have good experience designing **cloud**-based solutions in **Azure** by creating **Azure** **SQL** **database**, setting up Elastic pool jobs and designing tabular models in **Azure** analysis **services**. Created complex stored procedure, Views, Triggers, Functions, using PL/**SQL**, T-**SQL**.
* Strong knowledge in working with **ETL** methods for **data** extraction, transformation and loading in corporate-wide **ETL** Solutions and **Data** Warehouse tools for reporting and **data** analysis.
* Experience in all phases of **Data** Warehouse development like requirements gathering, design, development, implementation, testing, and documentation. Extensively used **Spark** **Data** Frames API over Cloudera platform to perform analytics on **Hive** **data**.
* Acquired profound knowledge in developing production ready **Spark** applications utilizing **Spark** Core, **Spark** Streaming, **Spark** **SQL**, **Data** Frames, Datasets and **Spark**-ML. Experienced and Implemented SAN TR migrations like Host based and Array based migrations.
* Expertise in building PySpark and **Spark** applications for interactive analysis, batch processing and stream processing
* Experienced in configuring **Git** and creating CICD pipelines and deployment tasks in **Azure** DevOps.
* Hands on experience in using **Hadoop** ecosystem components like **Hadoop**, **Hive**, Pig, **Sqoop**, HBase, Cassandra, **Spark**, **Spark** Streaming, **Spark** **SQL**, Oozie, Zookeeper, **Kafka**, Airflow, Flume, MapReduce framework, Yarn.
* StrongHadoopandplatformsupportexperiencewithalltheentiresuiteoftoolsandservicesinmajor **Hadoop** Distributions – Cloudera, **Azure** HDInsight, **AWS** and Hortonworks. Hands on experience in building applications using **python** libraries and packages.
* SkilledinusingAzureauthenticationandauthorizationandexperienceinusingVisualizationstoolslikeTableau, Power BI. Basic hands-on experience working with Kusto. Experience writing complex **SQL** queries, **SQL** Tuning and knowledge of **database** design and development (**SQL** and TSQL)
* Strong knowledge in working with **ETL** methods for **data** extraction, transformation and loading in corporate-wide **ETL** Solutions and **Data** Warehouse tools for reporting and **data** analysis. Good experience with Unit and Integration Testing process.
* WorkingknowledgewithAmazonS3, **Amazon** EC2, **AWS** Kinesis to provide a complete solution for computing, query processing and storage across a wide range of applications. Hands-on experience in building applications in **Hadoop** eco system.
* Good understanding of **Spark** Architecture, MPP Architecture, including **Spark** Core, **Spark** **SQL**, **Data** Frames, Driver Node, Worker Node, Stages, Executors and Tasks. Develop **data** visualizations, dashboards and reports and using Tableau and Power BI.
* Good experience in developing **Spark** applications using PySpark, **Spark** - **SQL**, and Delta Lake in **Databricks** for ingesting and transforming the **data**. Strong **Data** warehousing skills including **Data** profiling, **Data** masking, Auditing in **ETL**.
* Strong knowledge and experience in **data** modeling, **database** design, optimization for performance, stored procedures.
* Good knowledge in applying appropriate test methodologies including writing test plans and test cases.

**Education Qualification**

* Bachelors of Commerce, Osmania University, India-2017

**Technical Skill Set:**

|  |  |
| --- | --- |
| **Microsoft Azure** | **Azure** Databricks, **Azure** **Data** Factory, Synapse Analytics, HDInsight, ADLS, **Azure** Storage, **Data** Explorer, **Azure** Functions, Event Hub, Iot Hub, Logic apps, Stream Analytics, **Azure** **Web** App, **Azure** Analysis **Services**, Application Insights, **Azure** Active Directory, Key vault |
| **Hadoop/BigData/AWS Technologies** | Apache **Hadoop** 2.x/1.x, Cloudera CDP, Hortonworks HDP, **HDFS**, Map Reduce, **Sqoop**, **Hive**, Oozie, **Spark**, Zookeeper, **Kafka**, Airflow, Flume. **Amazon** **AWS**[ EMR, EC2, EBS, RDS, S3, Kinesis] |
| **Scripting Languages** | **Python**, **SQL**, **Scala**, **Java**, R, Pig, HiveQL, C,**C++** |
| **NOSQL Database** | **Azure** CosmosDB, MongoDB, DynamoDB, HBase, Cassandra |
| **Relational Databases** | **Azure** **SQL** **Database**, **SQL** Server, Oracle, MYSQL,PL/**SQL** |
| **Monitoring/Reporting** | PowerBI, Tableau, **Azure** Monitor, Log Analytics, Custom shell scripts |
| **Java Technologies** | Servlets, **JSP**, JDBC, EJB, Struts, Spring, Hibernate |
| **Web Development** | JavaScript, **HTML**, CSS, jQuery |
| **Development/Build Tools** | PyCharm,Anaconda,Eclipse,Ant,**Maven**,IntelliJ,Junit,log4J,Gradle |
| **Operating Systems** | Linux,Unix,MacOS-X,CentOS,Windows10,Windows8,Windows7 |
| **Version Control** | **GIT**, Bitbucket, SVN |
| **Methodologies** | Agile, Waterfall |

**Technical EXPERIENCE**

**Client: Corebridge Financial,Houston, Texas Duration: October 2021 - Present**

**Role:  Sr. Azure Data Engineer**

* Proficient in working with **Azure** **cloud** platform (**Data** Bricks, **Data** Factory, HDInsight, DataLake, Blob Storage, Synapse Analytics, **Azure** **SQL**, **SQL** pool, **Azure** Serverless apps). Conducted **data** analysis to identify trends and insights for business decision making.
* Designed, developed, and maintained **data** warehouse, **Data** Lake and **ETL** processes in **Azure** environment using **Azure** Synapse, **Azure** **Data** Lake, **and Azure** **Data** Factory, **Azure** Logic Apps and Synapse Dedicated **SQL** pool.
* Developed and maintained real-time and batch **data** ingestion processes using Apache Druid's ingestion APIs and Tranquillity ingestion service.
* Designed and implemented efficient Druid **data** schemas to support complex analytical queries on large-scale datasets.
* Tuned and optimized Druid cluster configurations for query performance and resource utilization, including segment size, indexing, and caching settings. An understanding of Snowflake's utilities, SnowSQL, SnowPipe, and the use of **Python** for **Big** **Data** models.
* Built and deployed custom Druid extensions, such as custom aggregators, filters, and functions, to support advanced analytical use cases.
* Integrated Druid with other **data** processing frameworks, such as Apache **Spark** and Apache **Kafka**, to enable **data** pipelines that combine real-time and batch **data**. Worked with cross-functional teams to understand business requirements and implement solutions accordingly.
* Implemented **data** modeling methodologies and techniques including ER Modeling, Dimensional Modeling to design Conceptual, Logical and Physical **Data** models. Developed and optimized **SQL** queries for performance tuning of **data** warehouse.
* Implemented DevOps practices for efficient code management and deployment. Experience in building the CI/CD pipelines and managing the injection of **data**. Experience in **data** processing like collecting, aggregating from various sources using Apache **Kafka** & Flume
* Dynamic **Snowflake** queries automated with dbt, **git**, and **Azure** DevOps, allowing for continuous integration and continuous delivery.
* Involved in building an Enterprise DataLake using **Data** Factory and Blob storage, enabling other teams to work with more complex scenarios and ML solutions. Generated various graphical planning reports using **Python** packages like NumPy, matplotlib.
* Wrote **SQL** Queries against **Snowflake** and Created **ETL** pipelines in and out of **data** warehouse using combination of **Python** and snowflakes SnowSQL. Utilized capabilities of Tableau such as **Data** extracts, **Data** blending, Forecasting, Dashboard actions and Table calculations.
* Designing and developing **Azure** **Data** Factory (ADF) extensively for ingesting **data** from different source systems like relational and non-relational to meet business functional requirements. Moved **data** from **Azure** storage to **snowflake database**.
* Contributed to the Apache Druid open-source community by submitting bug reports, patches, and new features, and participating in community discussions and events. Developed Streaming pipelines using **Azure** Event Hubs and Stream Analytics to analyze **data**.
* Created, provisioned multiple **Databricks** clusters needed for batch and continuous streaming **data** processing and installed the required libraries for the clusters. Developed **Databricks ETL** pipelines using notebooks, **Spark** **Data** frames, **SPARK** **SQL** and **python** scripting.
* Developed a **data** pipeline using **Kafka** and **spark** streaming to store **data** in **HDFS** and performed real time analytics on the incoming **data**.
* Experience in real time importing **data** to **Hadoop** using **Kafka** and implemented Oozie workflow for scheduled imports.
* Worked with various streaming ingest **services** with Batch and Real-time processing using **Spark** streaming, **Kafka** Confluent, Storm, Flume and **Sqoop**. Implement One time **Data** Migration of Multistate level **data** from **SQL** server to **Snowflake** by using **Python** and SnowSQL.
* Have good experience in setting up separate applications and reporting **data** tiers across servers using Geo replication functionality and failover groups. Extensively used **data** bricks notebooks for **data** processing and interactive analytics using **Spark** API’s
* Extensive knowledge in **data** transformations, mapping, cleansing, monitoring, debugging, performance tuning and trouble-shooting clusters.
* Designed custom-built input adapters using **Spark**, **Hive**, and **Sqoop** to ingest and analyze **data** (Snowflake, MS **SQL**, MongoDB) into **HDFS**.
* Experience in development of enterprise solutions using **Hadoop** components like **Hadoop** (**HDFS**, MapReduce, Yarn), **Spark**, SparkSQL, PySpark, **Kafka**, **Sqoop**, HBase, Impala, Pig, Flume, Oozie, Nifi, **Kafka**, Zookeeper, Storm, and **Hive**. Implement Copy activity, Custom **Azure** **Data** Factory Pipeline Activities. Collaborate with application architects on infrastructure as a service (IaaS) applications to Platform as a Service (PaaS).
* Extensively used **Azure** Active Directory, Role based Access Control (RBAC), Access Control Lists (ACL), created app registrations on **Azure** AD, Shared Access Signatures (**SAS**). Build the Logical and Physical **data** model for **snowflake** as per the changes required
* Performed validation and verified pipeline at all testing phases which includes Functional Testing, System Integration Testing, End to End Testing, Regression Testing, Sanity Testing, User Acceptance Testing, Smoke Testing, Disaster Recovery Testing, Production Acceptance Testing and Pre-prod Testing phases. Integrated Custom Visuals based on business requirements using Power BI desktop.
* Outlined custom-built input connectors utilizing Start, **Hive**, and **Sqoop** to ingest and analyze information (Snowflake, MS **SQL**, MongoDB) into **HDFS**. Worked with report analysts/or the Epic build teams to identify and resolve possible discrepancies between Cache and Clarity.
* For development, aggregation from different file formats such as XML, JSON, CSV, Avro, Parquet, and ORC, I've used **Python** (PySpark), **Java**, **Scala**, and **Spark**-**SQL**. Expert in developing JSON Scripts for deploying the Pipeline in **Azure** **Data** Factory (ADF) that process the **data**.
* Expert in using **Databricks** with **Azure** **Data** Factory (ADF) to compute large volumes of **data**. Expertise in **snowflake** to create and Maintain Tables and views. Developed an automated process in **Azure** **cloud** which can ingest **data** daily from **web** service and load in to **Azure** **SQL** DB.
* Performed **ETL** operations in **Azure** **Databricks** by connecting to different relational **database** source systems using jdbc connectors.
* Developed **Python** scripts to do file validations in **Databricks** and automated the process using ADF. Analyzed **data** where it lives by Mounting **Azure** **Data** Lake and Blob to Databricks. Used Logic App to take decisional actions based on the workflow.
* Developed custom alerts using **Azure** **Data** Factory, SQLDB and Logic App. Used **Python** and Shell scripts to Automate Teradata ELT and Admin activities. Worked on reporting tools OBIEE/BI publisher/PowerBI efficient **data** retrieval techniques and different types of reports
* Performed Application-level DBA activities creating tables, indexes, and monitored and tuned Teradata BETQ scripts using Teradata Visual Explain utility. Performance tuning, monitoring, UNIX shell scripting, and physical and logical **database** design.
* Developed UNIX scripts to automate different tasks involved as part of loading process. Worked on Tableau software for the reporting needs.
* Worked on creating few Tableau dashboard reports, Heat map charts and supported numerous dashboards, pie charts and heat map charts that were built on Teradata **database**. Design and implement **database** solutions in **Azure** **SQL** **Data** Warehouse, **Azure** **SQL**.
* Create pipelines in ADF using linked **services** to extract, transform and load **data** from multiple sources like **Azure** **SQL**, Blob storage and **Azure** **SQL** **Data** warehouse. Creating storage accounts which involved with end-to-end environment for running jobs.
* Design **data** auditing and **data** masking for security purpose. Monitoring end to end integration using **Azure** monitor.
* Utilized capabilities of Tableau such as **Data** extracts, **Data** blending, Forecasting, Dashboard actions and Table calculations.
* Implementation of **data** movements from on-premises to **cloud** in **Azure**. Involved in migrating the client **data** warehouse architecture from on-premises into **Azure** **cloud**. Developed **Sqoop** jobs for **data** ingestion, incremental **data** loads from RDBMS to **HDFS**.
* Developed **Azure** **Databricks** notebooks for handling large volumes of **data**, transformations, and computations to operate with various file formats. Implement **Azure** **Data** Factory operations and deployment into **Azure** for moving **data** from on-premises into **cloud**.
* Utilized **Spark**’s in memory capabilities to handle large datasets on S3 **Data** Lake. Loaded **data** into S3 buckets, then filtered and loaded into **Hive** external tables. Involved in developing a custom-built Rest API to support real time customer analytics for **data** scientists and applications.

**Environment: Azure Data Factory, Tableau, Shell Scripting, Teradata, python scripting, Azure data bricks, Azure data lake storage, Blob storage, Azure SQL Database, Azure synapse analytics, Azure synapse workspace, Synapse sql pool**,

**Client: Qwik Power Industries ,Hyderabad,India Duration: October 2018- May 2021**

**Role:  Hadoop Data Engineer**

* Installed and configured Flume, **Hive**, Pig, and Oozie on the **Hadoop** cluster.
* Collected and aggregated large amounts of **web** log **data** from different sources such as webservers, mobile and network devices using Apache Flume and stored the **data** into **HDFS** for analysis. Also used **Spark** **SQL** to handle structured **data** in **Hive**.
* Involved in making **Hive** tables, stacking information, composing **hive** inquiries, producing segments and basins for enhancement.
* Handled importing of **data** from various **data** sources, performed transformations using **Hive**, MapReduce, loaded **data** into **HDFS** and extracted **data** from Teradata into **HDFS** using **Sqoop**. Written and Implemented Teradata Fast load, Multi load and Bteq scripts, DML and DDL.
* Involved in migrating tables from RDBMS into **Hive** tables using **Sqoop** and later generate particular visualizations using Tableau.
* Analyzed substantial **data** sets by running **Hive** queries and Pig scripts. 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**.
* Defined the Accumulo tables and loaded **data** into tables for near real-time **data** reports.
* Created the **Hive** external tables using Accumulo connector. Ability to spin up different **AWS** instances including EC2-classic and EC2-VPC using **cloud** formation templates. Imported the **data** from different sources like **AWS** S3, LFS into **Spark** RDD.
* Involved in creating Shell scripts to simplify the execution of all other scripts (Pig, **Hive**, Impala and MapReduce) and move the **data** inside and outside of **HDFS**. Creating files and tuned the **SQL** queries in **Hive** utilizing HUE. Experienced in writing **Spark** applications in **Python** (PySpark).
* Expertise in implementing **Spark** using **Scala** and **Spark** **SQL** for faster testing and processing of **data** responsible to manage **data** from different sources. Worked with NoSQL databases like HBase in making the tables to load expansive arrangements of semi structured **data**.
* Designed the **ETL** process and created the high-level design document including the logical **data** flows, source **data** extraction process, the **database** staging, job scheduling and Error Handling Processed and transferred the **data** from **Kafka** into **HDFS** through **Spark** Streaming APIs.
* Created **ETL** Mapping with Talend Integration Suite to pull **data** from Source, apply transformations, and load **data** into target **database**.
* Working knowledge on creating **Kafka** topics, partitions and writing custom practitioner’s classes. Design relational and non-relational **data** stores on **Azure** Created Indicator, Pie Chart, Bar chart, and Scatter map to show different widgets on Sisense dashboard.
* Made use of the **AWS** environment to launch the applications in different regions and implemented CloudFront with **AWS** Lambda to reduce latency. Worked and acquired a lot of knowledge in **Amazon** **Web** **Services** (**AWS**) **cloud** **services** like EC2, S3, EMR, EBS, RDS and VPC.
* Extracting real time **data** using **Kafka** and **Spark** streaming by Creating D streams and converting them into RDD, processing it and storing it into Cassandra. Configured, deployed and maintained multi-node Dev and Test **Kafka** Clusters. Creating **ETL** test scenarios and test cases and plans to execute test cases. Involved in import and export of **data** between **Hadoop** **Data** Lake and Relational Systems like MySQL, Oracle using **Sqoop**. Sent requests to Confidential REST Based API from a **python** script via **Kafka** producer
* Experience in building Real-time **Data** Pipelines with **Kafka** Connect and **Spark** Streaming. Building the Cassandra nodes using **AWS** &amp; setting up the Cassandra cluster using Ansible automation tools. Written storm topology to emit **data** into Cassandra DB.
* Developed Oozie Bundles to Schedule Pig, **Sqoop** and **Hive** jobs to create **data** pipelines. Designed and developed **ETL** workflows and datasets in Alteryx. Developed **Hive** queries to do analysis of the **data** and to generate the end reports to be used by business users.
* Developed **Scala** scripts, UDF’s using both **Data** frames/**SQL**, RDD/MapReduce in **Spark** for **Data** Aggregation, queries and writing **data** back into RDBMS through **Sqoop**. Involved in the process of Cassandra **data** modelling and building efficient **data** structures.
* Develop batch processing solutions by using **Data** Factory and **Azure** **Data** bricks Implement **Azure** **Data** bricks clusters, notebooks, jobs and auto scaling. Design for **data** auditing and **data** masking Design for **data** encryption for **data** at rest and in transit
* Preparing **ETL** test strategy, designs and test plans to execute test cases for **ETL** and BI systems. Interacting with business users and understanding their requirements. Good understanding of **data** warehouse concepts. Good exposure and understanding of **Hadoop** Ecosystem
* Proficient in **SQL** and other relational databases. Good exposure to Microsoft Power BI. Good understanding and working knowledge of **Python** language. Experienced in working with **spark** eco system using **Spark** **SQL** and **Scala** queries on different formats like Text file, CSV file.
* Created a pipeline to gather **data** using PySpark, **Kafka** and Hbase Sent requests to source REST Based API from a **Scala** script via **Kafka** producer Created a pipeline to gather new music releases of a country for a given week using PySpark, **Kafka** and **Hive**
* Utilized a cluster of multiple **Kafka** brokers to handle replication needs and allow for fault tolerance Hands - on experience with **Spark** Core, **Spark** Session, SparkSQL, and **Data** Frames/**Data** Sets/RDD API **Spark** jobs, **Spark** **SQL** and **Data** Frames API to load structured **data** into **Spark** clusters Worked on importing and exporting **data** between the **HDFS** and Relational **Data** Bases
* Created a **Kafka** broker which uses the schema to fetch structured **data** in structured streaming Define **Spark** **data** schema and set up development environment inside the cluster Decoded the raw **data** and loaded into JSON before sending the batched streaming file over the **Kafka** producer Established a connection between the HBase and **Spark** for the transfer of the newly populated **data** frame
* Management of **Spark**-submit jobs to all environments Interacted with **data** residing in **HDFS** using PySpark to process the **data**
* Formatted the response into a **data** frame using a schema containing, News Type, Article Type, Word Count, and News Snippet to parse the JSON Design **Spark** **Scala** job to consume information from S3 Buckets Received the JSON response in **Kafka** consumer written in **Python**.
* Monitor background operations in Hortonworks Ambari **HDFS** Monitoring job status and life of the **Data** Nodes according to the specs
* Managed Zookeeper configurations and Z Nodes to ensure High Availability on the **Hadoop** Cluster Managed **hive** connection with tables, databases and external tables Setup ELK collections to all environments and replications of shards
* Create standardized documents for company all usage Work one on one with clients to resolve issues regarding **Spark** jobs submissions
* Worked on **AWS** Kinesis for processing huge amounts of real-time **data** Developed multiple **Spark** Streaming and batch **Spark** jobs using **Java**, **Scala**, and **Python** on **AWS** Involved in the creation of **Hive** tables, loading **data** from different sources and performing **Hive** queries.
* Implemented **data** queries and transformations on **Hive**/**SQL** tables by using the **Spark** **SQL** and **Spark** **Data** Frames APIs
* Experience collecting log **data** from various sources and integrating it into **HDFS** using Flume; staging **data** in **HDFS** for further analysis.
* Received the JSON response in **Kafka** consumer **python** file, formatted the response into a **data** frame using a schema containing, country code, artist name, number of plays and genre to parse the JSON Extracted metadata from **Hive** tables with **Hive** QL
* Established a connection between **Hive** and **Spark** for the transfer of the newly populated **data** frame Utilized a cluster of three **Kafka** brokers to handle replication needs and allow for fault tolerance Stored the **data** pulled from the API into Apache **Hive** on Hortonworks Sandbox

**Environment: Hadoop, Cloudera, HDFS, MapReduce, YARN, Hive, Pig, Sqoop, HBase, Apache Spark, Accumulo, Oozie Scheduler, Kerberos, AWS, Tableau, Sisense 4.7, Talend, HUE, HCATALOG, Flume, Git.**

**Client:  Kaalmanns Plastics Ltd ,Hyderabad,India Duration: August 2016 - September 2018**

**Role:  Data Engineer**

* Worked as a **Data** **Engineer**, analysed large **data** sets and provided efficient business solutions as per the requirements given by the Product Owner. Created both external and internal tables in **Hive**. Worked on Dynamic Partitioning. Worked intensively on **Spark** **Data** frames and **hive** tables. Loaded the **data** into **spark** **data** frames and performed necessary transformations and saved the results into **hive**.
* Developed **ETL** jobs using PySpark in which the **data** has been transformed in multiple stages and actions like aggregations are performed.
* Parsed Semi Structured JSON **data** and converted to Parquet using **Data** Frames in Pyspark.
* Fine Tuning and Optimizing the **Spark** Jobs using different configurations, cache and broadcast joins.
* Understand the Client requirements and designed the best possible approach to meet the Customer Use case.
* Involved in complete **Software Development Life Cycle (SDLC)** process by analyzing business requirements and understanding the functional workflow from source systems to target systems. Developed various **data** ingestion and **data** processing pipelines using **python**.
* Experienced in designing and developing highly effective **ETL** pipelines using multiple **big** **data** technologies. In - depth understanding of **Spark** Architecture including **Spark** Core, **Spark** **SQL**, **Data** Frames. Experience in managing and reviewing **Hadoop** log files.
* Expertise in various phases of project life cycles (Design, Analysis, Implementation and testing).
* Very Good experience working in **Azure** **Cloud**, **Azure** DevOps, **Azure** **Data** Factory, **Azure** **Data** Lake Storage, **Azure** Synapse Analytics, **Azure** Analytical **services** and **Data** bricks. Developed **spark** application by using **python** (py **spark**) to transform **data** according to business rules.
* Experience in building **data** pipelines using **Azure** **Data** Factory to get the **data** from source systems by using different **Azure** Activities like Move &Transform, Copy, filter, for each, **Databricks** etc. Experienced in working with different file formats like CSV, JSON, Avro and Parquet.
* Creating **Spark** clusters and configuring high concurrency clusters using **Azure** **Databricks** to speed up the preparation of high-quality **data**.
* Responsible for estimating the cluster size, monitoring and troubleshooting the **spark** **data** bricks cluster.
* Experience in Developing **Spark** applications using **Spark** - **SQL** in **Databricks** for **data** extraction, transformation and aggregation from multiple file formats for analyzing & transforming the **data**. Responsible for the maintenance of secure PII, CCPA **data** and **data** transfer.
* Experienced in performance tuning of **Spark** Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* Managed **data** coming from different sources and involved in **HDFS** maintenance and loading of structured and unstructured **data**
* Hands-on experience in handling **database** issues and connections with **SQL** databases by installing and configuring various packages in **python**. Created Dataframes and RDD’s using **spark** by reading the **data** from **Hive** Tables and parquet files
* Developed various **Python** scripts to find vulnerabilities with **SQL** Queries, permission checks and performance analysis and developed scripts to migrate **data**. Experience in using **GIT** Version Control System. Worked on **AWS** **Data** Pipeline to configure **data** loads from S3 to into **Redshift**
* Worked on complete **Software Development Life Cycle using Agile and Kanban Methodologies** to maintain the user stories and tasks.
* Involved in daily scrum meetings to discuss the development/progress and was active in making scrum meetings more productive.
* Work closely with development, test, documentation and product management teams to deliver high quality products and **services** in a fast-paced environment identified critical failure points during production and created test cases to address failure points.
* Participated in meetings with stakeholders and product managers to obtain domain knowledge and expectations of product
* Good Understanding of **AWS** **services** like **EMR, RDS, S3, Athena and EC2** instances. Knowledge on handling **Hive** queries using **Spark** **SQL** that integrates with **Spark** environment. Imported **data** from **AWS** S3 into **Spark** RDD, Performed transformations and actions on RDD's.
* Hands-on experience with **Amazon** **EC2, Amazon S3**, **Amazon** **RDS, IAM**, **Amazon** **Elastic Load Balancing**, and other **services** of the **AWS** family. Export the analyzed **data** to all the relational databases using **Sqoop** for visualization and to generate reports.
* Responsible for continuous monitoring and managing Elastic MapReduce cluster through **AWS** console.
* Worked on various execution enhancements like using distributed cache for small datasets, Partition, Bucketing in **Hive** and Map Side joins.
* Developed a Shell Script and **Python** script for running **Sqoop** commands, exception Handling and storing the logs while ingesting the Transactional and Electronic Invoice **data** from Oracle to **HDFS**.

**Environment: Hadoop, Map Reduce, Python, HDFS, Spark, GitHub, Microsoft Azure (Databricks, Azure Devops, Data Lake, Blob Storage, Azure Data Factory, Azure Active Directory, Azure Analytical services)**

**Client: Nile Ltd - Hyderabad, India Duration: February 2015 - July 2016**

**Role:  ETL Developer**

* Maintain and support the Teradata architectural environment for EDW Applications. Interacted with the business community and gathered requirements based on changing needs. Performed Teradata and Informatica tuning to improve the performance of the Load.
* Designed the logical **Data** Model using Erwin and transformed the Logical model to Physical **database** using Power Designer. **Data** Mart and Dimensional Modeling, Star and **SnowFlake** Schema Modeling for **Data** Warehouse.
* Developed mappings/scripts to extract **data** from Oracle, Flat files, **SQL** Server, DB2 and load into **data** warehouse using the Mapping Designer, BTEQ, Fast Load and Multi Load. Exported **data** from Teradata **database** using Fast Export and BTEQ.
* Wrote **SQL** Queries, Triggers, Procedures, Macros, Packages and Shell Scripts to apply and maintain the Business Rules. Coded and implemented PL/**SQL** packages to perform batch job scheduling. Performed error handling using error tables and log files.
* Assisted the team to work on the installation and configuration of databases. Used Table auto create dashboards based for management.
* Responsible for maintaining **SQL** server databases and performing **data** validation for complex **SQL** queries.
* Used **python** and **SQL** for developing **ETL** pipelines and loaded the transformed **data** into **SQL** Server.
* Participated in the development and maintenance of company **data** security, privacy, policies, procedures, and best practices.
* Performed code reviews and assisted developers in optimization and troubleshooting. Wrote **UNIX** **shell scripts** for file handling and scheduling.
* Resolved bugs in the pipelines and addressed queries from end users. Updated **data** dictionary for all published **data** assets.
* Implemented and maintained **data** analysis scripts using **SQL** and **Python**. Ingested **data** from disparate **data** sources and created **data** views to be used in Tableau. Worked with various file formats such as Parquet, Avro, and various compression formats like Snappy.
* Wrote **SQL** queries to perform backend testing of the front-end application. Performed unit testing and coordinated with programmers to fix them. Processed multiple terabytes of **data** stored in **AWS** using **Elastic Map Reduce (EMR)** to **AWS** **Redshift**
* Created the dimensional **data** warehouse model design, developed the **ETL** jobs loading the disposition codes tables from Pivotal into the **Data** Warehouse. Created stored procedures to **ETL** **data** from multiple sources into the **data** warehouse,
* Designed and created the supporting **ETL** design documents. Wrote shell scripts to automate ingestion tasks.
* Used Cron jobs to schedule the execution of data processing scripts.

**Environment: SQL Server 2012, Tableau, T-SQL, Python, Windows, UNIX**