**Rohan Mehta | Big Data Engineer**

**Employer Details:-**

**Parker@stiersol.com** **PHN: (610) 553 6336**

**PROFESSIONAL SUMMARY:**

* Around 5 years of work experience in Data Engineering and implementation of Hadoop, Spark and cloud Data Warehousing solutions.
* Experience working with core Apache Hadoop components and tools like HDFS, Map Reduce, Hive, Hive QL, HBase, Spark, Flume, Sqoop, Oozie, Pig, Cassandra etc.
* Extensive experience with Python, Scala and Shell languages.
* Experience in Spark eco-system, core SQL, and streaming modules**.**
* Hands-on experience with core Apache Hadoop architecture components such as HDFS, MapReduce, Name nodes, Data nodes, Secondary Name node, YARN etc.
* Good experience with Amazon Web Services (AWS) Big data services – S3, EC2, ECS, ECR, EMR Spark, Redshift, Glue, Kinesis etc.
* Good experience in Snowflake data warehouse, developed data extraction queries, automatic ETL for data loading from Data Lake.
* Experience working on Redshift Database which includes creating clusters and backups and creating Star-schema, data modeling and copying data from S3 buckets and performing ETL jobs on Glue.
* Worked on Glue ETL pipelines from S3 to Redshift (batch jobs).
* Worked on AWS step functions for creating workflows for batch jobs on AWS.
* Good experience working with Athena on querying data on S3 buckets by partitioning (year to min).
* Developed Sqoop scripts for incremental data transfer between Hadoop and RDBMS.
* Experience in developing Spark applications using core Spark, Spark-SQL, and Spark Streaming API’s with Scala and Python.
* Experience in integrating NoSQL databases technologies including HBase, Apache Cassandra and MongoDB.
* Experience in working with different operating systems such as MS Windows, MacOS, UNIX, Linux – Ubuntu and CentOS.
* Experience in source code & build management with Git & Enterprise GitHub with Jenkins.
* Involved in optimizing Snowflake cost optimization initiatives with data models & efficient queries.
* Strong understanding of real time streaming technologies Spark and Kafka.
* Worked on Cloudera Distribution (CDH) and Hortonworks Data Platform (HDP).
* Possess excellent communication, interpersonal and analytical skills along with a positive attitude.
* Experience in dealing with Apache Hadoop components like HDFS, MapReduce, HiveQL, HBase, Pig, Sqoop, Big Data and Big Data Analytics.
* Experience in building scripts using Maven, Gradle and using CI/CD tools like Jenkins.
* Expertise in using Kafka as a messaging system to implement real-time streaming solutions and implementing Sqoop for large data transfers from RDMS to HDFS/HBase/Hive and vice-versa.
* Hands on experience in developing ETL jobs in Hadoop eco-system using Oozie & NiFi, streaming and batch jobs in AWS ECS & Lambda, EMR Spark and docker containers.
* Hands on experience in installing, configuring and using components like Hadoop, MapReduce, HDFS, HBase, Zookeeper, Hive, Sqoop and Pig with Hortonworks Data Platform.

**EDUCATION:**

Master of Science in Technology – Computer Certificate, Eastern Illinois University, USA

Bachelor of Technology, Rai Univ., India

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
|  **Programming/Scripting Languages** | Python, Scala, SQL, Java, Shell |
| **Big Data, Streaming** | Hadoop, MapReduce, HDFS, Hive, Pig, Sqoop, Zookeeper, Oozie, YARN, Spark 2.0, Kafka, Kinesis Firehose & Data Streams, SQS, SNS, Athena. |
| **AWS** | EC2, S3, ECS, ECR, Lambda, ELB, Redshift, Athena, Glue, EMR, Kinesis, MSK |
| **Databases** | Oracle, DB2, SQL Server, NoSQL, Apache Cassandra, HBase, Snowflake |
| **Cloud Computing** | Amazon Web Services (AWS) |
| **Other tools & OS** | Microsoft Office tools, VMware, IoT, Git, Enterprise GitHub, PyCharm, and Maven, Slack, MS Windows, UNIX, Linux, MacOS |
| **File Formats** | Text, XML, JSON, Avro, Parquet, ORC, Protobuf |
| **DevOps** | Ansible, Jenkins, Git, GitHub |
| **Methodologies** | Agile, Waterfall, UML, Design Patterns |

**PROFESSIONAL EXPERIENCE:**

**Client: Credit Suisse, Raleigh, NC**

**Role: Data Engineer Jun 19 – Present**

Project: Cardholder Transaction History (CTH)

Description: Latest data attributes changes in compliance & discard the deprecated columns from HBase tables. New attributes have been added to specific column families. Identification of charge backs & fraud detection in MapReduce modules has been enhanced according to compliance parameters. A one-time reusable batch process has been designed to apply compliance changes in CTH data, it is being utilized for every year consecutively.

**Responsibilities:**

* Worked with Hadoop 2.x version and Spark 2.x (Python and Scala).
* Used Spark for interactive queries, processing of streaming data and integration with NoSQL database for huge volume of data.
* Experienced in handling large datasets using Partitions, Spark in-memory capabilities, Broadcasts in Spark, Effective &amp; efficient Joins, Transformations and other during ingestion process itself.
* Developed custom ETL solutions, batch processing and real-time data ingestion pipeline to move data in and out of Hadoop using Python and shell scripting.
* Explored with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark SQL, DataFrame, Pair RDD and Spark on YARN.
* Worked with Sqoop import and export functionalities to handle large data set transfer between Oracle databases and HDFS.
* Creating S3 buckets and managing policies for S3 buckets and utilized S3 buckets and Glacier for archival storage and backup on AWS.
* Experience with AWS Lambda workflow implementation using python to interact with application deployed on EC2 instance and S3 bucket.
* Developed Spark jobs to clean data obtained from various feeds to make it suitable for ingestion into Hive tables for analysis.
* Built an AWS-EC2 instance and Migrated data to Cloud.
* Developed Custom Input Formats in Spark jobs to handle custom file formats.
* Configured Oozie workflow to run multiple Hive jobs which run independently with time and data availability.
* Utilized Hive tables and HQL queries for daily and weekly reports. Worked on complex data types in Hive like Structs and Maps.
* Developed HiveQL queries for trend analysis and pattern recognition on user data.
* Helped this regional bank streamline business processes by developing, installing and configuring Hadoop ecosystem components that moved data from individual servers to HDFS.
* Spark streaming is used for a single framework to satisfy all their processing needs.
* Imported data from AWS S3 into Spark RDD, performed transformations and actions on RDDs.
* Worked with Hadoop distribution of Hortonworks.
* Containerize and deploy critical jobs that run more than Lambda runtime in AWS ECS.
* Perform data quality check jobs that runs in Python & Spark on Snowflake data, multiple parallel jobs for each query and aggregating data from all query results and verifying data quality according to defined standards.
* Data quality job designs different data plot in graphical way and sends email to responsible stakeholders every day.
* Created HBase tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, NoSQL and a variety of portfolios.
* Supported code/design analysis, strategy development and project planning.
* Created reports for the BI team using Sqoop to export data into HDFS and Hive.
* Developed multiple Spark jobs in Scala & Python for data cleaning and preprocessing.
* Assisted with data capacity planning and node forecasting.
* Collaborated with the infrastructure, network, database, application and BI teams to ensure data quality and availability.
* Designing ETL processes using Informatica to load data from Flat Files, Oracle and Excel files to target Oracle Data Warehouse database.

**Environment:**  Spark, Spark SQL, Spark Streaming, HBase, Hive, Oozie, Informatica, HQL, NoSQL, MySQL, RDBMS, Oracle DB, Sqoop, Scala, Python, Shell scripting, AWS, EC2, ECS, ECR, S3 buckets, EMR, Glue Maven, GIT, Tableau, Core Hadoop.

**Client: Cigna, Bloomfield, CT**

**Role: Big Data Engineer Oct 18 – Jun 19**

**Project: Data Warehouse Migration**

**Description**: Implementing Data Warehouse migration from traditional MySQL data base to Hadoop Eco system. Large structural tables from MySQL to HDFS and create Star-schema on Hive tables. Validations against schema and Sqoop import from MySQL and design and implementation of Oozie workflows for all tables-based requirements.

**Responsibilities:**

* Participated in the creation, distribution and walkthrough of software test cases, scripts and other documents surrounding testing activities and ensure that all testing activities and deliverables are conducted/ produced in compliance with company standards.
* Worked under the Cloudera distribution.
* Involved in working with Sqoop for fetching the data from RDBMS.
* Transformed and stored the ingested data into data frames using spark SQL
* Created Hive tables to load the transformed Data.
* Performed partitions and bucketing in hive for easy data classification.
* Worked on Performance and Tuning optimization of Hive.
* Involved in exporting Spark SQL Data frames into hive tables stored as ORC Files.
* Involved in Ingesting real-time log data from various producers using Kafka.
* Used spark streaming to subscribe to desired topics (drug) for real time processing.
* Transformed the dstreams into data frames using spark engine.
* Experienced in performance tuning of Spark Application for setting right Batch Interval time, level of parallelism and memory tuning for optimal efficiency.
* Performed data cleansing to meet business requirements.
* Responsible for performing sort, join, aggregations, filter, and other transformations on the data.
* Appended the data frames to pre-existing data in hive.
* Performed analysis on the hive tables based on the business logic.
* Created a data pipeline using oozie workflows which performs jobs on a daily basis.
* Involved in Analyzing data by writing queries in HiveQL for faster data processing.
* Involved in Persisting Metadata into HDFS for further data processing.
* Loading data from Linux File systems to HDFS and vice-versa using shell commands.
* Used Git as Version Control System.

**Environment:** Spark, AWS, EC2, EMR, Hive, SQL Workbench, Maven, Kibana, Sqoop, Spark SQL, Spark Streaming, Scala, Python, Hadoop (Cloudera Stack), Hue, Spark, Netezza, Kafka, HBase, HDFS, Hive.

**Client: Thomson Reuters Corporation**

**Role: Big Data Developer Jan 18 – Oct 18**

**Responsibilities:**

* Involved in review of functional and non-functional requirements.
* Utilized SQOOP, Flume and Hadoop File System API’s for implementing data ingestion pipelines.
* Worked on data streaming using Flume, performed transformations on the data using Pig & Hive.
* Worked on importing metadata into Hive and migrating existing tables and applications to work on Hive.
* Utilized SQOOP, Flume and Hadoop File System API’s for implementing data ingestion pipelines.
* Imported data using Sqoop to load data from MySQL to HDFS on regular basis.
* Developing Scripts and Batch Job to schedule various Hadoop Program.
* Written Hive queries for data analysis to meet the business requirements.
* Creating Hive tables and working on them using Hive QL.
* Importing and exporting data into HDFS and Hive using Sqoop.
* Got good experience with NOSQL database i.e. HBase.
* Involved in creating Hive tables loading with data and writing hive queries which will run internally in map reduce way.
* Developed a custom File system plug in for Hadoop so it can access files on Data Platform.
* Setup and benchmarked Hadoop/HBase clusters for internal use.

**Environment:** Hadoop, HDFS, Hive, Hortonworks distribution, Spark, CDH, Flume Linux, XML, MySQL, Hadoop, HDFS, ETL, YARN, Pig, Hive.

**Client: Polaris, Hyderabad, India**

**Role: Hadoop Developer Jul 16 – Jul 17**

**Responsibilities:**

* Responsible for building scalable distributed data solutions using Hadoop.
* Building and sizing the Hadoop cluster based on the data extracted from all the sources.
* Monitored Hadoop cluster job performance and capacity planning.
* Responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring and troubleshooting, manage and review data backups, manage and review Hadoop log files.
* Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.
* Involved in loading data from LINUX file system to HDFS.
* Importing the data from RDBMS to HDFS using Sqoop.
* Developed Simple to complex Map Reduce Jobs to transform the ingested data.
* Optimized Map Reduce Jobs to use HDFS efficiently by using various compression mechanisms.
* Created Hive tables to load the transformed Data.
* Designed and Implemented Partitioning, Bucketing in HIVE for better classification.
* Performed different joins on Hive tables and implementing Hive SerDe's like REGEX, JSON.
* Performed transformations using Hive QL, Map Reduce, and loaded data into HDFS.
* Developed hive queries to analyze/transform the data.
* Used Oozie workflow engine to automate Map Reduce jobs.

**Environment:** Hadoop, MapReduce, HDFS, Hive, Java, Maven, Hadoop distribution of Horton Works, Cloudera, HBase, Linux, XML, MySQL, Hadoop, ETL, Kafka, YARN, Hive, Cassandra.

**Client: IGATE - Hyderabad, India**

**Role: Hadoop Developer Jan 16 – Jul 16**

**Description:** The project is for developing a web-based application to eliminate all the paperwork in the laboratories, reading the data from different instruments and store the data in a relational database and generating business intelligence reports as per the requirement for the management.

**Responsibilities:**

* Involved in schema validation for MySQL to Hive tables.
* Utilized SQOOP, Hadoop File System API’s for implementing data ingestion pipelines.
* Worked on Hive SQL, performed transformations on the data using Hive
* Worked on importing metadata into Hive and migrating existing tables and applications to work on Hive.
* Implanting Data extraction and store in HDFS along with 3 replications.
* Imported data using Sqoop to load data from MySQL to HDFS on regular basis using Oozie.
* Developing Scripts and Batch Job to schedule various Hadoop Program.
* Written Hive queries for table creation and data analysis to meet the business requirements.
* Creating Hive tables and working on them using Hive QL.
* Importing and exporting data into HDFS and Hive using Sqoop.
* Got good experience with NOSQL database i.e. HBase.
* Involved in creating Hive tables loading with data and writing hive queries which will run internally in map reduce way.
* Worked with JSON and Parquets file formats on imports and exports in Hive for Hadoop so it can access files on Data Platform.

**Environment:** Hadoop, HDFS, Hive, Cloudera Distribution (CDH), Hortonworks (HDP), Linux, XML, MySQL, Hadoop, HDFS, ETL, YARN, Hive.