Shiwangi Bhatia

+91 8168376988
sbhatia2193@gmail.com

Sr. Data Engineer with 4 years of experience in driving the data-driven solutions to increase efficiency, accuracy and creating data solutions, and analyzing data structure and Size to deliver insights and implement action-oriented solutions to complex business problems. Extensive hands on experience in **Big Data technologies like Map Reduce,Apache NIFI, Spark, Hive, Pig, NoSQL etc**. Have good functional and technical knowledge . **Experience of cloud technologies i.e AWS.**

Experience Summary

- Working on configuring Data ingest with **Apache SQOOP,Apache NIFI(v1.10)** in to the target HDFS and Cassandra
- Spark Architecture and Components : **Spark Core & PySpark SQL with Dataframes**
- Big Data Technologies: **Hadoop(CDH Distribution),** MapReduce framework & Spark Ecosystem.
- POC experience **with Kubernetes and Docker Swarm** for containerized solution.
- Involved in technology migration from SAS to Python and Spark.
- Delivered data analysis projects using hadoop based tools and the python data science stack on top of AWS(S3 & EMR)
- **Developed job flows in Apache Nifi allowing data flow and extraction , system failure & Scheduling.**
- 2.5yrs Hive joins, performance tuning in joins.
- 2yr Sqoop tools for data ingestion, import data into Hadoop ecosystem.
- Python Scripting for Data Structure Transformation and Automation
- POC experience on AWS - S3,EMR,Glue,ATHENA,SNOWFLAKE,DMS,Lambda willing to take up new work challenges on AWS.
- PySpark MLib POC with Regression Models & Classifiers.

Highlights

Data Engineering Stack :

- **Data Ingestion** :- Sqoop,Apache NiFi(v1.10),Apache Pig
- **NoSQL DB** :- HBase,Cassandra
- **Frameworks** :- Hadoop (HDFS & MapReduce), Apache Spark(2.X)
- **Programming**- SQL, Scala, Python, Core Java
- **Cloud** - AWS
- **Others** :- kafka, Streaming,Chronos Scheduler

Education

Bachelor of Technology (IT)- JMIT,Radaur(YNR) (2012 – 2016)

Under Graduation/10+2: 84% From GRM, Delhi, India

Activities/Interest

- Cooking
- Reading Spiritual Context
- Travel
- Great food

Projects Overview : -

1. **Nagarro,Senior Data Engineer**
   Sep 2019 – Present(Domain :- Telcom)

---------- **Reporting Solution for Telcom Client(Jan 2018 – Present)** ----------

Environment: Apache Spark(2.x),SBT , Python,Cassandra(3.x), Apache NIFI(1.10v)

Architechture Layers : Apache Nifi (Ingestion Layer) || Processing Layer Apache Spark(Scala) || Supporting Frameworks (HDFS, YARN, MapReduce, Cassandra, Spark, Apache Nifi, Zookeeper, Oozie , Yarn , Spark SQL,Spark Streaming)

Team Size : 10

Description : Client is is Australia's largest mobile network that provides users with mobile phones, internet plans and packages, home phones & more.The project is about ingestion data of various CDR and ingestion of same into Cassandra in order to generate Buisness Reports on the top of same after processing in Spark

Role and Responsibilities :

- Development of Ingestion Pipeline in order to load data into Cassandra using Apache NiFi.
- Development of Spark Jobs using Intellij and Integration of the same with GIT.
- Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and Driver/Executor Memory tuning and Optimizations required as per data size and Joins
- Flattening JSON Data in order to filter and Update on the basis of nested Values using JOLT Tranformations
- After Insertion of RAW Data into Cassandra , post that applying various rejects & validations based on business logic in Spark , In the third layer we are applying SCD type 2 & some of our own transformations.
- During the fourth layer we use one year & data of quarter of different dimensions & facts from staging layer to create segmentations & finally publishing the data to PostgresDB

----------**European Telcom Client(*Sept* 2019 – Dec 2019)** ----------

Environment: Apache Spark(2.x), Python ,HDFS, Streamsets,ElasticSearch,Kibana,Kafka

Architechture Layers : Streamsets || Processing Layer Apache Spark(Scala) || Supporting Frameworks (HDFS, YARN, MapReduce, Hive, Spark, Zookeeper,)

Team Size : 5

Description : The primary objective of this project is to create a Ingestion Pipeline is to load two types of data (Batch and Streaming) .Batch data is loaded directly into HDFS and Streaming data is loaded using Kafka Streamsets Connector.

Role and Responsibilities :

- Development of Ingestion Pipeline in order to load data into HDFS and create streaming data using Kafka Producer and Consumer Processors using Apache Streamsets.
- Development of Spark Jobs using Intellij and Integration of the same with GIT.
- Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and Driver/Executor Memory tuning and Optimizations required as per data size and Joins
- Optimizing of existing algorithms in Hive using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
- Insertion of data for ouput spark job into Elasticsearch in order to create Visualization and Dashboards On the top of the same in Kibana.

2. Accenture,Senior Data Engineer
March – August 2019

---------- **RBS Client: Athena Replatforming(March 2018 – August 2019)**----------

Environment: AWS(EMR cluster), S3 for Storage, Apache Pyspark for ETL processing(Pyspark Dataframes and Spark SQL)

Description: Project is about Converting Existing Data Model from SAS to Pyspark and loading SAS extracts to S3 Buckets and using EMR cluster on AWS .Using Athena for faster querying .

Responsibilities:

- Responsible for coding existing SAS logic for 200 trench 1 tables in Pyspark and Spark SQL.
- Ingested SAS extracts from SAS server to AWS.
- Dynamic creation of column names using data frame logics using latest year month values
- Implementing merge logic of SAS using joins in Pyspark
- Implementation of clustered tables in Pyspark using Indexing .
- Loading month wise tables into Hive using partitions on yaer/month values.
- Using Thread abstraction to create concurrent threads of execution.

3. TCS , System Engineer
October 2016 – Feb 2018

---------- **RX-Offload- Hadoop Developer** ----------

Environment: Hadoop, Map-Reduce, Pig, Hive, Java, Big-Data, Sqoop, Oracle, Teradata, Datastage.

Description: Project is about Converting Existing Data Model from Teradata, and Data Stage to Hadoop using different Hadoop components.

Modules:

- Load Ready Merge
- Load Ready
- Noise Reduction
- Raw Ingestion(Hadoop Scripts)

Responsibilities:

- Responsible for designing, coding, testing of New Model in Hadoop.
- Ingested file from a third server to main server and then to HDFS using a Raw Ingestion Framework.
- Built Hive scripts and did reformation with actual data Using Pig and finally inserted data back to Hive tables with dynamic partitioning.
- Assisted In designing Data Quality framework through which we can handle ^M like character in hadoop as in hadoop these character behave as Line breaker by default.
- Built Sqoop for transferring of data from hive tables to Teradata and vice-versa.
- Used special ORC Format tables with UTF-8 encoding so that all the special character are handled perfectly even in Hive tables.
- Developed Pig and hive scripts which will load the data into final hive tables
- Converting JDBC code written for Teradata to HBASE Java API code and implementing the filter classes in java and comparison of results using Soap UI
- Successfully implemented Data Compaction logic using Apache Hive and Pig.

---------- **Banking Client - AML Data Ingestion** ----------

Environment: Hadoop, Map-Reduce, Hive, Java, Big Data, Sqoop, Oracle

Description: AML (Anti Money Laundering)- DI(Data ingestion) team is responsible to ingest AML application data to Cornerstone (Big data platform) which is currently hosted in IDN (Teradata Platform) applying the required transformation on existing data.

Responsibilities:

- Creation of metadata of existing data/schema
- Analysing various business and transformation requirements and coming up with logic to handle changes in existing data structure/schema
- Working with the various source teams which are sending data to AML to understand the existing data/schema and their relevance.
- Writing the Hive queries and scripts to ingest data.
- Ingesting data using internally created modules
- Unit testing of ingested data and validation of transformations happened.

---------- **Declaration** ----------

I hereby declare that the above particulars are true & correct to the best of my knowledge and belief. In the event of any Information being found false or incorrect, my candidature will be liable to be canceled.

Place: Gurugram

Date: 14-07-2020