**PROFESSIONAL EXCELLENCE SUMMARY:**

* 8+ years of IT industry experience exploring various technologies, tools and databases like **Big Data, AWS, S3, Snowflake, Hadoop, Hive, Spark, python, Sqoop, CDL(Cassandra), Teradata, Tableau, SQL, PLSQL** and **Redshift**.
* Having **5+ years of experience as Big Data Application Designer** (following **agile methodology**) and **2 years of experience as PL/SQL developer.**
* Having 5+ years of experience in **End-to-end in Big Data Application Designing** with strong experience on major components of Hadoop Ecosystem like **Apache** **Hadoop Map Reduce, Strom, Spark, HDFS, HIVE, PIG, HBase, Zookeeper, Sqoop, Oozie, Kafka, Python, Scala.**
* 4+ Years of Experience in **NoSQL databases HBase & MongoDB** and **search** **engines like Solr and Elastic search.**
* Good hands-on experience in algorithm development in map-reduce and Spark.
* Created **Hive** tables, dynamic partitions, buckets for sampling and working on them using **Hive QL**.
* Experience in Analysing and implementing data mining algorithm including customized key, value type and custom file formats
* Experience in managing and reviewing **Hadoop** log files.
* Managing the Hadoop distribution with Cloudera Manager, Cloudera Navigator, and **Hue**.
* Extensive experience in developing applications using **JSP, Servlets, Spring, Hibernate, Java Script, Angular, AJAX, CSS, HTML, JDBC, JNDI, JMS, XML,** and **SQL across the platforms like Windows, Linux, and UNIX.**
* Extensive experience in building batch and steaming data pipelines using cutting edge technologies (**Docker, Kubernetes, Hadoop, AWS and AZURE**).
* Good Hands-on experience on **Python, SQL,** and **R.**
* Good understanding of **RDD** operations in **Apache Spark** like Transformations &Actions, Persistence/ Caching, Accumulators, Broadcast Variables, Optimizing Broadcasts.
* Hands on experience in performing aggregations on data using **Hive Query Language (HQL).**
* Good experience in extending the core functionality of **Hive** and **Pig** by developing user-defined functions to provide custom capabilities to these languages.
* Expertise in Hadoop Ecosystem components **HDFS, Map Reduce, Hive, Pig, Sqoop, HBase and Flume for Data Analytics.**
* Demonstrated capacity to implement innovative solutions that enhances customer satisfactions, decreases cost, enhances efficiency, eliminates downtime, and strengthen organizations.
* Worked in hive-HBase integration and retrieving **HBase** data from the hive and vice versa.
* Good Knowledge in Amazon Web Service (**AWS**) concepts like EMR and EC2 web services which provide fast and efficient processing of Teradata Big Data Analytics and using S3 as a storage mechanism.
* Involved in developing Dashboards using **Tableau** and P**ower BI** to generate the reports.
* Hands on experience on Data Analytics Services such as **Athena, Glue Data Catalog & Quick Sig**
* Very good in Application Development and Maintenance of SDLC projects using different programming languages such as **Java, C, Scala, SQL, and NoSQL**.
* Experience in optimization of Map reduce algorithm using combiners and partitioners to deliver best results.
* Experience in analyzing and implementing data mining algorithm including customized key, value type and custom file formats
* In-depth understanding of Data Structure and Algorithms.
* Hands-on experience in handling database issues and connections with SQL and NoSQL databases like **MongoDB, Cassandra, Redis, CouchDB, DynamoDB** by installing and configuring various packages in **python.**
* Experience in developing Custom UDFs for datasets in Pig and Hive.
* Analyse latest Big Data Analytic technologies and their innovative applications in both business intelligence analysis and new service offerings.
* Designed and Developed Shell Scripts and **Sqoop** Scripts to migrate data in and out of **HDFS**
* Designed and Developed **Oozie** workflows to execute **MapReduce** jobs, Hive scripts, shell scripts and sending email notifications
* Deployed different partitioning methods like Hash by field, Round Robin, Entire, Modulus, and Range for bulk data loading
* Hands on experience in working with input file formats like parquet, json, Avro.
* Worked on Extraction, Transformation, and Loading (**ETL**) of data from multiple sources like Flat files, XML files and Databases.
* Used Agile Development Methodology and Scrum for the development process.

**Technical Summary:**

|  |  |
| --- | --- |
| Hadoop/ Big Data | Apache Spark, Hadoop, HDFS, Map Reduce, PIG, Hive, Sqoop, Oozie, Flume, HBase, YARN, Cassandra, Phoenix, Airflow |
| Frameworks | Hibernate, Spring, Cloudera CDHs, Hortonworks HDPs, MAPR |
| Programming & Scripting Languages | Java, Python, R, C, C++, HTML, JavaScript, XML, Git |
| Database | Oracle 10g/11g, PostgreSQL, DB2, SQL Server, MySQL, Redshift |
| NoSQL Database | HBase, Cassandra, MongoDB |
| IDE | Eclipse, Net beans, Maven, STS (Spring Tool Suite), Jupiter Notebook |
| ETL Tools | Pentaho, Informatica, Talend |
| Reporting Tool | Tableau, Power BI |
| Operating Systems | Windows, UNIX, Linux, Sun Solaris |
| Testing Tools | Junit, MR Unit |
| AWS | EMR, Glue, Athena, Dynamo DB, Redshift, RDS, Data Pipelines, Lake formation, S3, IAM, CloudFormation, EC2, ELB/CLB. |
| Azure | Data Lakes, Data Factory, SQL Data warehouse, Data Lake Analytics, Databricks, other azure services. |

**Client: Charles Schwab**

**Location: Westlake, TX   
Role: Sr. Data Engineer Nov 2021 – Present**

**Responsibilities:**

* **Anchor artifacts** for multiple milestone (**application design, code development, testing, and deployment**) in software lifecycle.
* **Develop Apache Strom** program to consume the Alarms in **real time streaming** from Kafka and enrich the alarm and pass it to EEIM Application.
* **Creating rules Engine in Apache Strom** to categorize the alarm into Detection, Interrogation & Association types before processing of alarms.
* **Develop Apache Spark** program using python language (**PYSPARK**) to establish the establish a connection between **Mongo DB** and EEIM application.
* Responsible to develop **EEIM Application as Apache Maven project** and commit to code to **GIT**.
* Analyse the Alarms and enhance the EEIM Application using **Apache Strom to predict the root cause** of the alarm and exact device where the network failure is happened.
* Accumulate the EEIM Alarm data to the **NoSQL database called Mongo DB** and retrieve it from Mongo DB when necessary.
* Build Fiber To The Neighbourhood or Node (FTTN) Topology and Fiber To The Premises (FTTP) Topology using **Apache Spark and Apache Hive**.
* Categorize the real time streaming Alarms into Matched Alarm, Unmatched Alarm and Unparsed alarm and store it into **Apache Solr search engine**.
* **Review the performance of the system** and revaluate the platform by doing the complete system regression test with heavy load of data and capture the logs and metrics of performance review
* Process the system logs using **Logstash tool** and store to elastic search and **create dashboard using Kibana.**
* Regularly **tune performance of Hive** queries to improve data processing and retrieving
* Create statistical reports in **Banana dashboard** using the data in Apache Solr search.
* Provide the technical support for **debugging, code fix, platform issues, missing data points**, unreliable data source connections and big data transit issues.
* Developed **Java and Python application** to call the external **REST APIs to retrieve weather, traffic, geocode information.**
* Worked on TAX POC from scratch using **Sqoop and hive**.
* Review the **unit, integration, system, regression test** results of Data Pipelines in the development environment and provide GO or NO GO for the system to Production.
* Conducting **code reviews on regular basis or on ad-hoc / on-demand** when AT&T deems necessary.
* Creation of **simulation tools and data sets** for unit and integration testing.
* Provide the advanced simplistic approaches by researching the data using **Machine Learning and Deep Leaning Techniques.**
* Provide analytics on most failed equipment in topology using the **H2O analytical tool** and build a dashboard.
* Experienced with **Jira, Bit Bucket** and source control systems like **GIT and SVN and** development tools like **Jenkins, antifactory.**

**Environment:** HDFS, Hive, Pig, Sqoop, Spark, Scala, Python, MapReduce, Hortonworks, Teradata, Zookeeper, MySQL, Shell Scripting, Informatica, Ubuntu, Linux Red Hat, GitHub, Edge node.

**Client: Epam  
location: Florida**

**Role: Data Engineer May 2019 - Oct 2021**

**Responsibilities:**

* Assisted in upgrading, configuration and maintenance of various Hadoop infrastructures like **Pig**, **Hive,** and **HBase**.
* Configured **Flume** to capture the news from various sources for testing the classifier.
* 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** using Azure Data factory.
* Analyse, design, and build Modern data solutions using Azure PaaS service to support visualization of data. Understand current Production state of application and determine the impact of new implementation on existing business processes.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL Azure DW) and processing the data in In Azure Databricks.
* Experience in Developing **Spark** applications using **Spark** - **SQL** in **Databricks** for data extraction, transformation, and aggregation from multiple file formats for analysing & 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.
* Experience in developing **MapReduce** jobs using various Input and output formats.
* Assisted in upgrading, configuration, and maintenance of various **Hadoop** infrastructures like **Pig**, **Hive**, and **HBase**.
* Developed workflow in **Oozie** to automate the tasks of loading the data into **HDFS** and pre-processing, analyzing, and training the classifier using **MapReduce jobs, Pig jobs** and **Hive jobs**.
* Developed **Spark** scripts by using **Scala shell** commands as per the requirement.
* Involved in loading data into **Cassandra** NoSQL Database.
* Developed **Spark** applications to move data into **Cassandra** tables from various sources like Relational Database or Hive.
* Worked on **Spark streaming** collects the data from **Kafka** in near real time and performs necessary transformations and aggregations on the fly to build the common learner data model and persists the data in **Cassandra**.
* Developed Hadoop Solutions on **AWS** from Developer to Admin roles utilizing the **Hortonworks Hadoop Stack.**
* Managed **RHL/AWS** Role Based Security and **Hadoop Admin Load Balancing** on **AWS EC2 Cluster.**
* **Migration Solution** from **on prem** **to AWS** using **Sqoop, Pig, AWS Cloud**
* Worked on **Cassandra** Data modelling, NoSQL Architecture, DSE **Cassandra** Database administration, **Key space creation**, **Table creation**, **Secondary** and Solr Jenkin index creation, User creation & access administration.
* Experience in performance tuning a **Cassandra** cluster to optimize **writes** and **reads**.
* Developed **Python** scripts, **UDF'**s using both **Data** frames/**SQL** and **RDD**/**MapReduce** in **Spark** for **Data Aggregation**, queries and writing data back into RDBMS through Sqoop.
* Used **Pig** and **Hive** in the analysis of data.
* Load the data into **Spark** **RDD** and performed in-memory data computation to generate the output response.
* Developed **Spark** code using **Scala** and Spark-SQL/Streaming for faster testing and processing of data.
* Performed Sqooping for various file transfers through the **HBase** tables for processing of data to several **NoSQL DBs- Cassandra, MongoDB**.

**Environment:** Hadoop, Hive, HBase, Spark, Scala, GIT, Sqoop, Kafka, Cassandra, Python, Spark RDD, AWS, Pig, Cloudera, IntelliJ, Agile, and Jira

**Client: HMS Insurance   
location: Charlotte, NC**

**Role: Data Engineer Jan 2018 - April 2019**

**Responsibilities:**

* Handled importing of data from various data sources, performed transformations using **Hive**, **Map reduce**, loaded data into **HDFS**.
* Extracted the data from **MySQL** into **HDFS** using **Sqoop**.
* Exported the analyzed data to the **Relational databases** using **Sqoop** for visualization and to generate reports for the BI team.
* Developed Simple to complex **Map Reduce** jobs.
* Analyzed the data by performing **Hive** **queries** and running **Pig Scripts** to know user behavior.
* Created partitioned tables in **Hive**.
* Administered and supported distribution of **Horton works**.
* Wrote **Korn shell**, **Bash shell**, **Pearl scripts** to automate most **DB** maintenance tasks.
* Worked on Installed and configured **Hadoop Map Reduce**, **HDFS**, developed multiple **Map Reduce** jobs in **Java** for data cleaning and preprocessing.
* Importing and exporting data into **HDFS** and **HIVE** using **SQOOP.**
* Responsible to manage data coming from different sources.
* Monitoring the running **Map Reduce** programs on the cluster.
* Responsible for loading data from **UNIX** file systems to **HDFS**.
* Installed and configured **Hive** and Created **Hive** **UDF**s.
* Hands on experience with Apache Spark using Scala implemented spark solution to enable real time report from **Cassandra** data.
* Good Command on **CQL** to run queries on the data present in Cassandra Cluster with multi DC’s in 8 nodes each.
* Involved in creating **Hive Tables**, loading with data and Writing **Hive queries** which will invoke and run **Map Reduce** jobs in the backend.
* Implemented the **workflows** using **Apache Oozie** framework to automate tasks.
* Developed scripts and automated data management from end to end and sync up between the clusters.

**Environment:** Apache Hadoop, Java, Bash, ETL, Map Reduce, Hive, Pig, Horton works, Deployment tools, Data tax, Flat files, Oracle 11g/10g, MySQL, Window NT, UNIX, Sqoop, Oozie.

**Client: AT&T**

**Role: Data Engineer May 2016 - Dec 2017**

**Responsibilities:**

* Involve in all phases of **SDLC (Software Development Life Cycle),** including **requirement collection, design, analysis, development**, and **application deployment**.
* **Architecture** and **design** of business requirements and making **Visio Diagrams** for designing and developing the application and deploying the application in various environments.
* **Develop Spark 2.1/2.4 Scala** component to process the business logic and store the computation results of **10 TB data** into **HBase database** to access the downstream web apps using **Big** **SQL** **db2 database**.
* Uploaded and processed more than ten terabytes of data from various structured and unstructured sources into **HDFS** using **Sqoop and Flume**. Test the developed modules in the application using **Junit** Library and **Junit Testing Framework**
* Analyse structured, unstructured data, and file system data and load the data to **HBase** tables based on the project requirement using **IBM Big SQL** with **Sqoop mechanism** and processing the data using **Spark SQL** in-memory computation &processing results to **Hive**, **HBase**
* Handle integrating additional enterprise data into **HDFS** using **JDBC**, loading **Hadoop** in **Big SQL**, and performing transformations on the fly using **Spark API** for developing the standard learner data model, which obtains data from upstream in near real-time and persists into **HBase.**
* I am working with different file structures with other **Hive** file formats like **Text file**, **Sequence file**, **ORC file, Parquet**, and **Avro** to analyze the data to build a data model and read them from **HDFS** and process through parquet files and loading into **HBASE** tables.
* Develop the Batch jobs using **Scala programming language** to process the data from files and tables, transform the data with the business logic and deliver it to the user
* Work in the **Continuous Deployment module**, which is used to create new tables or update the existing table structure if needed in different environments, along with **DDL (Data Definition Language**) creation for the tables. Also, I wrote **AZURE POWERSHELL** scripts to copy or move data from the local file system to **HDFS Blob storage**.
* Created **Pipelines in ADF (Azure Data Factory**) using **Linked Services/Datasets/Pipeline/ to Extract, Transform and load data** from different sources like **Azure SQL, Blob storage, Azure SQL Data warehouse**, **write-back tool**, and **backward.**
* Developed **JSON** Scripts for deploying the Pipeline in **Azure Data Factory (ADF)** that processes the data.
* Loading data from **Linux/Unix** file system to **HDFS** and working with **PUTTY** for the better communication between **Unix** and **Window system** and for accessing the data files in the **Hadoop environment;**
* Developed and Implemented **HBase** capabilities for Big de-normalized data set and then applied transformation on the de-normalized data set using **Spark/Scala**.
* Involved **Spark** tuning to improve the Jobs performance based on the **Pepper Data** **monitoring tool metrics**. Worked in building application platforms in the Cloud by leveraging **Azure Databricks**
* Develop **shell scripts** for configuration checks and files transformation, which is to be done before loading the data into the **Hadoop** Landing area **HDFS**
* Developed and Implement **Spark ETL** custom component to extract the data from upstream systems and push the data to **HDFS** and finally store the data in **HBase** with wide row format Work with **Apache Hadoop** environment by **Hortonworks**
* Enhance the application with new features and make the performance improvement in all the modules of the application
* Exposure to **Microsoft Azure** in the processing of migrating the on-prem data to **azure** **Cloud**, studying and implementing **spark** techniques like partitioning the data with Keys and writing it to parquet files which boost speed improvement.
* Understand the **Mapping documents**, existing Source Data and preparing load strategies for different source systems, and implement them using **Hadoop technology**
* Worked with Continuous integration tools like **maven, Team city, IntelliJ** and scheduled the jobs with **TWS (Tivoli Workload Scheduler)** tool, Creating and cloning the jobs and Job streams in the **TWS tool** and promoting them to higher environments.
* I coordinated with co-developers, agile development and project management team, and external systems and was responsible for demos, presentation of developed modules to the **project management team**.
* Perform **Code review activities** with peer developers in the team and architect of the team for delivering exception/error-free application, testing it with real-time test scenarios, and deploying it to the next level environments. Analyse and fix the data processing issues and troubleshoot the process.

**Environment:** Scala, Java, Spark framework, Linux, Jira, Bitbucket, IBM Big SQL, Hive, HBase, IntelliJ IDEA, Maven, Db2 Visualizer, ETL, TeamCity, WinSCP, PuTTY, IBM TWS (Tivoli Workload Scheduler), Windows, Azure Data Factory, Linux.

**Client: Infosys**

**Location: Hyderabad, India**

**Role: Data Engineer Jan 2014 - Dec 2015**

**Responsibilities:**

* Involved in design and development phases of Software Development Life Cycle (SDLC)
* Involved in designing **UML** Use **case diagram**, **Class diagram**, **Sequence Diagrams** and **Rational Rose**.
* Building a revenue-generating java-based web application using **JAVA**/**J2EE** technologies.
* Participating on development as well as integration of and enhancements to existing products.
* Used **agile** methodology and **SCRUM** meeting to track, optimize and tailored features to client requirement.
* User help tooltips implemented with **Dojo** Tooltip Widget with multiple custom colors.
* Experience and Developed user interface using **JSP**, **JSP Tag** Libraries and **Java Script** to simplify the complexities of the application.
* Experience and implemented **Model View Controller** (MVC) architecture using **Jakarta** **Struts** Frameworks at presentation tier.
* Followed and Developed as **Dojo** based front end including forms and controls and programmed event handling.
* Implemented **SOA** architecture with web services using **JAX-RS** (REST) and **JAX-WS** (SOAP)
* Developed various Enterprise **Java Bean** components to fulfill the business functionality.
* Implemented and created **Action Classes** which route submittals to appropriate **EJB** components and render retrieved information.
* Participating on analysis, design, build, unit test, deployment and support of the systems.

**Environment:** Core Java, J2EE, Oracle, SQL, Server, JSP, Jenkins, Dojo, Struts, spring, JDK, Hibernate, JavaScript, HTML, CSS, AJAX, Junit, Web services.